**MKT 680 Marketing Analytics**

Report for Project 2: Recommender Systems

Group 7: Myra Liu, Khasim Shaik, Jacky Yang, Jingwen (Kivi) Zuo

Date: 02/28/2019

**Overview**

This report is about recommending 2 products from at most 5 suppliers to at least 500 selected customers who purchased more than $5,000 in 2017 from Pernalonga, a leading supermarket chain in Lunitunia. Pernalonga wants to experiment on personalized promotions funded by partnering with suppliers to grow revenue. The objectives are to target the most profitable and promotion-sensitive customers and to recommend products with the highest expected value. In addition, the metrics applied to define the target customers and products will be discussed in more details.

**Business & Data Understanding**

Firstly, some issues are identified to better prepare the data:

* Bags is the most frequently purchased item in the dataset, but bags should not be included to conduct the recommender system from the business understanding.
* The transaction ID is not unique for all transactions. Base on the assumption that a transaction is made when a customer visits a specific store on a given day, the unique identifier for a transaction should be a combination of customer ID, transaction ID and store ID.
* A dubious transaction identified in the dataset is that one transaction has a negative paid amount. Without further details about this transaction, it is removed.  
  *Product id: 357541011, cust\_id: 93409897 tran\_prod\_paid\_amt: -0.55*

Secondly, the focus of analysis in the project is based on transaction level, as the metrics created measure the impact of promoted products’ on transaction values. It is more valuable to find products that induce more profitable transactions for the most promotion-sensitive customers.

Thirdly, due to the fact that some products sold by the retailer do not belong to any supplier as “No Label” or belong to the retailer as “Private Label”, they do not satisfy the requirement of partnering with suppliers. They are removed from the dataset.

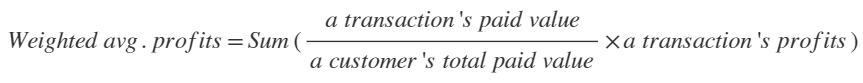
**Target Customer**

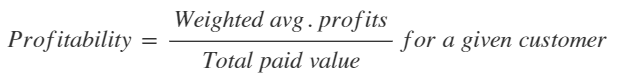
The complete data includes over 29 million historical transactions for 7,920 customers in 2016 and 2017. According to the scope of the project, there are two thresholds applied first, and there are 1,897 customers left in the dataset:

* Customers who purchased over L$5,000 in 2017
* The product that was purchased at least 15 times in 2017

In order to narrow down the scope and efficiently target a certain group of customers who are most likely to purchase more when they have promoted items in the transactions, two criteria are created to assess and filter the target customer. To clarify some of the major dollar amounts that we will use below, the sales amount equals to the sum of the discounted amounts and the paid amounts in the dataset.

* Profitability measures the profits generated in a transaction in a ratio of weighted average profits over the total paid amount generated on the customer level. Without the products’ cost information, the lowest discounted price of a given product will be used as a benchmark of cost to measure the same product’s profits in other transactions, as it is assumed that the lowest discounted price is higher than or equal to the actual cost. The importance factor used to calculate the weighted average profits for a customer is the ratio of a single transaction’s paid value over the customer’s total paid value, and the aggregated profitability is the sum of the product of the ratio and the profits generated over total paid amounts in the transaction.





The final profitability for a customer is measured on the transactional level, as it is important to identify the profits triggered by a promoted product. The process can be explained more clearly as below:

For each customer:

For each transaction:

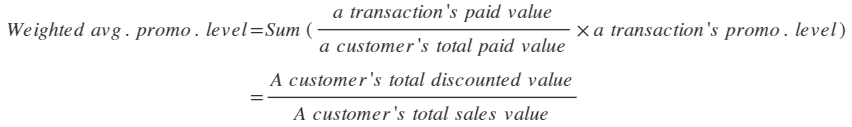
For each product:

Calculate the profits

Sum the profits for transactions with discounted items

Calculate the ratio of profits/total sales

* Promotion-sensitivity measures if a customer’s transaction value increases with higher promotion level, and it is defined by the correlation of transactions’ weighted average promotion levels and sales value for a given customer within a timeframe. The importance factor used to calculate weighted average promotion level is the same as for profitability, and the promotion level equals the transaction’s total discounted amount divided by total sales amount.



The process can be explained more clearly as below:

For each customer:

For each transaction:

For each products:

Calculate the discount level

Calculate the weighted average discount level on transaction level based on the product’s paid amount/transaction value

Conduct the correlation of the weighted average discount level vs. transaction value

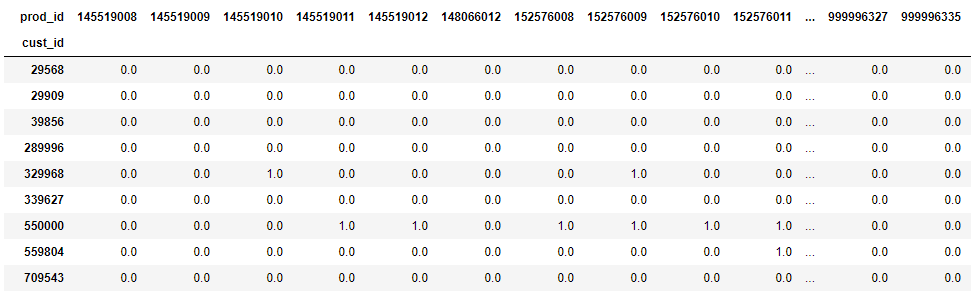
The final target customer is filtered by a threshold of 0.5 on both profitability and promo-sensitivity. Therefore, there are 1,348 customers left as the customer base for the recommender system.

**Similarity Measurement**

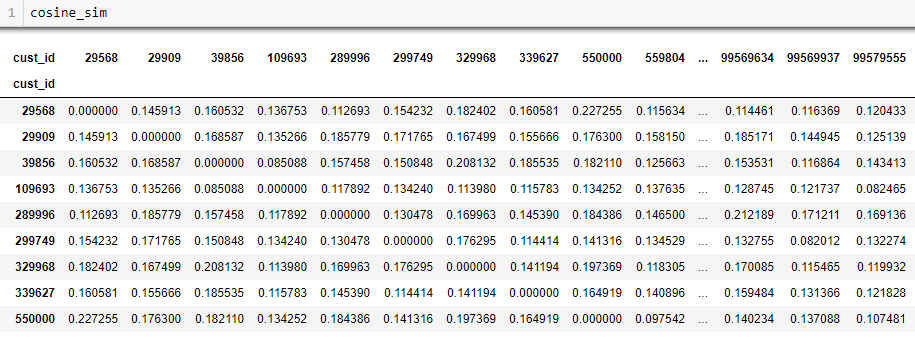
After narrowing down to the customer base, the next step is to calculate the similarity between all pairs of customers using their baskets of promoted products purchased in 2017. The input for the cosine similarity algorithm is a matrix of binary variable on only promoted products for each customer. We used cosine similarity rather than jaccard similarity because the similarity result will be the same when the “rating input” is binary. The binary matrix is used because of lack of rating information and the similarity is based on if the customer purchased a promoted product. The columns represent the product pool that includes all promoted products bought by the target customers. The rows represent the basket for each customer in vectors. Using only promoted products is to match with incremental sales that are calculated based on promoted products.

Example below:

Matrix of customer-product (1: purchased, 0: not purchased)



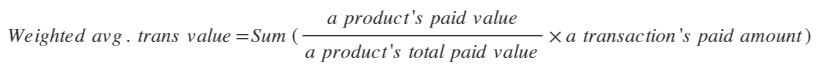
Matrix of similarity score



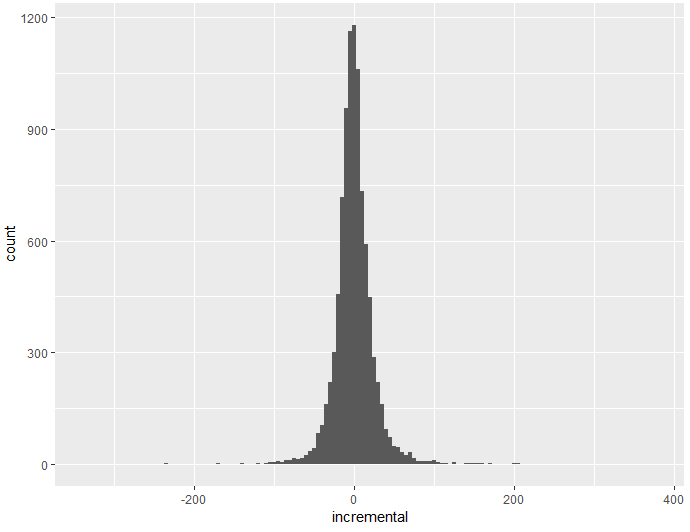
**Incremental Sales**

Before computing the purchase probability, incremental sales is also calculated to be implemented to the select the products that will be recommended to customers. Incremental sales measure the difference between a product’s weighted average transaction value with itself on promotion and without promotion on product level.

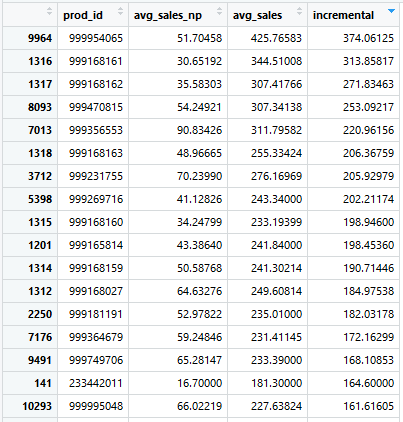
The transactions applied to the calculation is firstly filtered by the target customer base, as the incremental sales should be specifically targeting the selected customer base. The weighted average sales for each product is calculated as below with promotion and without promotion respectively:



The distribution of incremental sales for products purchased by the target customer is normally distributed with a mean of 0.1 and a maximum of $374.1.



A subset of top-ranked incremental sales is shown below:

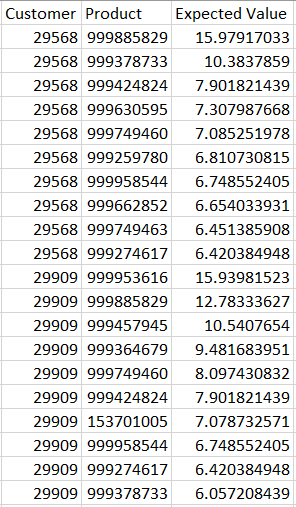


**Purchase Probability**

Next, the top 20 most similar customers are identified for each target customer by the ranking of similarity scores. A product pool is formed by all products purchased by this group of customers. The products that a customer has purchased are not eliminated because promotion of purchased products could also induce a larger transaction, which will be assessed by incremental sales in next step. Therefore, each product’s probability of purchase within the customer group could be calculated by the ratio of customers who purchased out of all 21 customers. As a result, each customer has a list probability of buying a product. The list of purchase probability is embedded in the codes and implemented to calculate expected value.

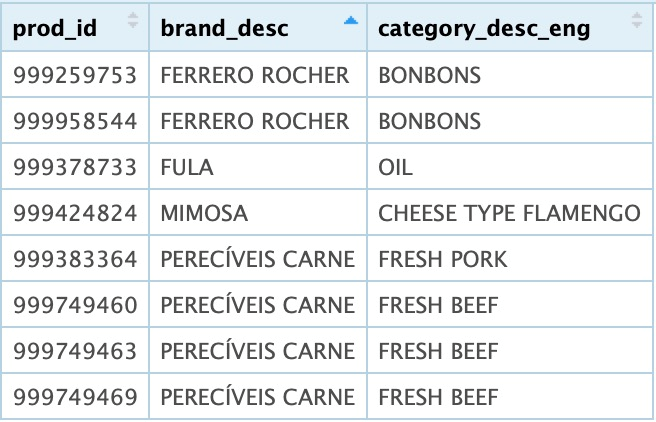
**Expected Value**

The expected value is measured by the product of purchase probability and incremental sales, and it is used to comprehensively rank the top products that will be recommended to customers. As the recommended products have to be limited within 5 suppliers, 5 top-ranked products are kept to better narrow down the supplier selection in the next step. A subset of 5 top-ranked products for customer 29568 and 29909 is shown below.



**Supplier Selection**

Due to the scope of the project, at the meantime of recommending two products to target customers, the suppliers should also be limited to at most five. A more comprehensive view is to create a table of customers, products and the product’s expected value for this customer. The total expected value will help with identifying the top 5 suppliers that will generate the highest expected value in the target customer base.



The final suppliers are decided as Ferrero Rocher, Fula, Mimosa and Pereciveis Carne by purchase frequency on promotion in the target customer base. Although Serrata ranks as the third in the list, it is removed from the list because only one oil supplier should kept and Fula generates higher expected value in the target customer base.The total expected value of the four suppliers is about $16,012 for the target customer base and an average of $20.63 from each customer.

**Final Recommended List**

The complete list can be referred to the “Recommendation\_list.csv” file.

